JC13 Rec'd PCT/PTO 03 APR 2001

FORM PTO-1390 (Modified) (REV 11-98) U.S. DEPARTMENT OF COMMERCE PATENT AND TRADEMARK OFFICE TRANSMITTAL LETTER TO THE UNITED STATES DESIGNATED/ELECTED OFFICE (DO/EO/US) CONCERNING A FILING UNDER 35 U.S.C. 371		ATTORNEY'S DOCKET NUMBER **MATR-0018-US** U.S. APPLICATION NO. (IF KNOWN, SEE 37 CFR 09/806756
INTERNATIONAL APPLICATION NO. **PCT/FR00/02220**	INTERNATIONAL FILING DATE **2 August 2000**	PRIORITY DATE CLAIMED **4 August 1999**

TITLE OF INVENTION

METHOD AND DEVICE FOR DETECTING VOICE ACTIVITY

APPLICANT(S) FOR DO/EO/US

STEPHANE LUBIARZ, EDOUARD HINARD, FRANCOIS CAPMAN and PHILIP LOCKWOOD

Applicant herewith submits to the United States Designated/Elected Office (DO/EO/US) the following items and other information:

1. ☒ This is a **FIRST** submission of items concerning a filing under 35 U.S.C. 371.
2. ☐ This is a **SECOND** or **SUBSEQUENT** submission of items concerning a filing under 35 U.S.C. 371.
3. ☒ This is an express request to begin national examination procedures (35 U.S.C. 371(f)) at any time rather than delay examination until the expiration of the applicable time limit set in 35 U.S.C. 371(b) and PCT Articles 22 and 39(1).
4. ☐ A proper Demand for International Preliminary Examination was made by the 19th month from the earliest claimed priority date.
5. ☒ A copy of the International Application as filed (35 U.S.C. 371 (c) (2))
   - a. ☒ is transmitted herewith (required only if not transmitted by the International Bureau).
   - b. ☐ has been transmitted by the International Bureau.
   - c. ☐ is not required, as the application was filed in the United States Receiving Office (RO/US).
6. ☒ A translation of the International Application into English (35 U.S.C. 371(c)(2)).
7. ☒ A copy of the International Search Report (PCT/ISA/210).
8. ☐ Amendments to the claims of the International Application under PCT Article 19 (35 U.S.C. 371 (c)(3))
   - a. ☐ are transmitted herewith (required only if not transmitted by the International Bureau).
   - b. ☐ have been transmitted by the International Bureau.
   - c. ☐ have not been made; however, the time limit for making such amendments has NOT expired.
   - d. ☐ have not been made and will not be made.
9. ☐ A translation of the amendments to the claims under PCT Article 19 (35 U.S.C. 371(c)(3)).
10. ☐ An oath or declaration of the inventor(s) (35 U.S.C. 371 (c)(4)).
11. ☐ A copy of the International Preliminary Examination Report (PCT/IPEA/409).
12. ☐ A translation of the annexes to the International Preliminary Examination Report under PCT Article 36 (35 U.S.C. 371 (c)(5)).

Items 13 to 20 below concern document(s) or information included:

13. ☐ An Information Disclosure Statement under 37 CFR 1.97 and 1.98.
14. ☐ An assignment document for recording. A separate cover sheet in compliance with 37 CFR 3.28 and 3.31 is included.
15. ☒ A **FIRST** preliminary amendment.
16. ☐ A **SECOND** or **SUBSEQUENT** preliminary amendment.
17. ☐ A substitute specification.
18. ☐ A change of power of attorney and/or address letter.
19. ☒ Certificate of Mailing by Express Mail
20. ☒ Other items or information:

A copy of the Request.

Express Mail No. EL732849804US

JC08 Rec'd PCT/PTO 03 APR 2001

U.S. APPLICATION NO. (IF KNOWN, SEE 37 CFR) 09/806756	INTERNATIONAL APPLICATION NO. PCT/FR00/02220	ATTORNEY'S DOCKET NUMBER MATR-0018-US

21. The following fees are submitted:.

BASIC NATIONAL FEE (37 CFR 1.492 (a) (1) - (5)) :

		CALCULATIONS	PTO USE ONLY
☐ Neither international preliminary examination fee (37 CFR 1.482) nor international search fee (37 CFR 1.445(a)(2) paid to USPTO and International Search Report not prepared by the EPO or JPO	**$1,000.00**		
☒ International preliminary examination fee (37 CFR 1.482) not paid to USPTO but Internation Search Report prepared by the EPO or JPO	**$860.00**		
☐ International preliminary examination fee (37 CFR 1.482) not paid to USPTO but international search fee (37 CFR 1.445(a)(2)) paid to USPTO	**$710.00**		
☐ International preliminary examination fee paid to USPTO (37 CFR 1.482) but all claims did not satisfy provisions of PCT Article 33(1)-(4)	**$690.00**		
☐ International preliminary examination fee paid to USPTO (37 CFR 1.482) and all claims satisfied provisions of PCT Article 33(1)-(4)	**$100.00**		
ENTER APPROPRIATE BASIC FEE AMOUNT =		**$860.00**	
Surcharge of **$130.00** for furnishing the oath or declaration later than ☒ 20 ☐ 30 months from the earliest claimed priority date (37 CFR 1.492 (e)).		**$130.00**	

CLAIMS	NUMBER FILED	NUMBER EXTRA	RATE		
Total claims	36 - 20 =	16	x $18.00	**$288.00**	
Independent claims	3 - 3 =	0	x $80.00	**$0.00**	
Multiple Dependent Claims **(check if applicable).**			☐	**$0.00**	
	TOTAL OF ABOVE CALCULATIONS =			**$1,278.00**	
Reduction of 1/2 for filing by small entity, if applicable. Verified Small Entity Statement must also be filed (Note 37 CFR 1.9, 1.27, 1.28) **(check if applicable).** ☐				**$0.00**	
	SUBTOTAL =			**$1,278.00**	
Processing fee of **$130.00** for furnishing the English translation later than ☐ 20 ☐ 30 months from the earliest claimed priority date (37 CFR 1.492 (f)). +				**$0.00**	
	TOTAL NATIONAL FEE =			**$1,278.00**	
Fee for recording the enclosed assignment (37 CFR 1.21(h)). The assignment must be accompanied by an appropriate cover sheet (37 CFR 3.28, 3.31) **(check if applicable).** ☐				**$0.00**	
	TOTAL FEES ENCLOSED =			**$1,278.00**	
				Amount to be: refunded	$
				charged	$

☒ A check in the amount of **$1,278.00** to cover the above fees is enclosed.

☐ Please charge my Deposit Account No. in the amount of to cover the above fees. A duplicate copy of this sheet is enclosed.

☒ The Commissioner is hereby authorized to charge any fees which may be required, or credit any overpayment to Deposit Account No. **20-1504.** A duplicate copy of this sheet is enclosed.

NOTE: Where an appropriate time limit under 37 CFR 1.494 or 1.495 has not been met, a petition to revive (37 CFR 1.137(a) or (b)) must be filed and granted to restore the application to pending status.

SEND ALL CORRESPONDENCE TO:

Dan C. Hu
TROP, PRUNER & HU, P.C.
8554 Katy Freeway, Suite 100
Houston, Texas 77024-1805
(713) 468-8880 [Phone]
(713) 468-8883 [Fax]

SIGNATURE

DAN C. HU
NAME

40,025
REGISTRATION NUMBER

April 3, 2001
DATE

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant: Stephane Lubiarz et al.	§	Group Art Unit:	
Serial No.: PCT/FR00/02220	§	Examiner:	
Filed: August 2, 2000	§		
For: METHOD AND DEVICE FOR DETECTING VOICE ACTIVITY	§	Atty. Dkt. No.:	MATR-0018-US

Box PCT
Commissioner for Patents
Washington DC 20231

PRELIMINARY AMENDMENT

Sir:

Prior to Examination, please amend the above-identified application as follows

In the Specification:

Page 1, at line 2, please insert the following paragraph:

--BACKGROUND OF THE INVENTION--

Page 2, at line 3, please insert the following paragraph:

--SUMMARY OF THE INVENTION--

Page 2, delete lines 36-37.

Page 3, delete lines 1-2.

Page 3, at line 3, please insert the following paragraph:

--BRIEF DESCRIPTION OF THE DRAWINGS--

Page 3, at line 25, insert the following paragraph:

--DETAILED DESCRIPTION--

Page 4, line 32, please replace the formula with the following:

$$S_{n,i} = \frac{1}{f(i) - f(i-1)} \sum_{f \in [f(i-1), f(i)]} S_{n,f}$$

Express Mail No. EL732849804US

Page 5, line 13, please replace the formula with the following (formula has not changed but replaced for clarity):

$$\hat{B}_{n,i} = \lambda_B . \hat{B}_{n-1,i} + (1-\lambda_B).S_{n,i}$$

Page 6, line 9, please replace the formula with the following (formula has not changed but replaced for clarity):

$$Hp_{n,i} = \frac{S_{n,i} - \alpha'_{n-\tau 1,i} . \hat{B}_{n-\tau 1,i}}{S_{n-\tau 2,i}}$$

Page 7, line 3, please replace the formula with the following (formula has not changed but replaced for clarity):

$$E_{1,n,j} = \sum_{i=i\min(j)}^{i\max(j)} [f(i) - f(i-1)] . \hat{E}p_{1,n,i}^{2}$$

Page 7, line 4, please replace the formula with the following (formula has not changed but replaced for clarity):

$$E_{2,n,j} = \sum_{i=i\min(j)}^{i\max(j)} [f(i) - f(i-1)] . \hat{E}p_{2,n,i}^{2}$$

Page 7, line 25, please replace the formula with the following (formula has not changed but replaced for clarity):

$$\overline{E}_{1,n,j} = \lambda . \overline{E}_{1,n-1,j} + (1-\lambda).E_{1,n,j}$$

Page 7, line 26, please replace the formula with the following (formula has not changed but replaced for clarity):

$$\overline{E}_{2,n,j} = \lambda . \overline{E}_{2,n-1,j} + (1-\lambda).E_{2,n,j}$$

Page 11, line 11, please replace the formula with the following (formula has not changed but replaced for clarity):

$$S'_{n,i} = \max\left(S_{n,i} - \alpha . \hat{B}_{n-1,i} ; \beta . \hat{B}_{n-1,i}\right)$$

Page 11, line 27, please replace the formula with the following (formula has not changed but replaced for clarity):

$$\hat{E}p_{n,i} = \max\left(S_{n,i} - f(\rho).\hat{B}_{n-1,i};\ \beta.\hat{B}_{n-1,i}\right)$$

<u>In the Claims:</u>

Cancel claims 14 and 15, without prejudice.

Amend the following claims:

1. (Amended) Method for detecting voice activity in a digital speech signal in at least one frequency band, wherein the voice activity is detected on the basis of an analysis comprising the step of comparing two different versions of the speech signal, wherein at least one of said versions is a denoised version obtained by taking account of estimates of noise included in the signal.

2. (Amended) Method according to claim 1, wherein said comparison is performed on respective energies, evaluated in said frequency band, of the two different versions of the speech signal, or to a monotonic function of said energies.

3. (Amended) Method according to claim 1, wherein said analysis further comprises a time smoothing of the energy of one of said versions of the speech signal, and a comparison between the energy of said version and the smooth energy.

4. (Amended) Method according to claim 3, wherein the comparison between the energy of said version and the smooth energy controls transitions of a voice activity detection automaton from a speech state to a silence state, and wherein the comparison of the two different versions of the speech signal controls transitions of the detection automaton from the silence state to the speech state.

5. (Amended) Method according to claim 1, wherein the two different versions of the speech signal are two versions denoised by non-linear spectral subtraction, wherein a first of the two versions is denoised in such a way as not to be less, in the spectral domain, than a first fraction of a long-term estimate representative of a noise component included in the speech signal, and the second of the two versions is denoised in such a way as

not to be less, in the spectral domain, than a second fraction of said long-term estimate, smaller than said first fraction.

6. (Amended) Method according to claim 5, wherein said analysis further comprises a time smoothing of the energy of each of the two versions of the speech signal, by means of a smoothing window determined by comparing the energy of the second of the two versions with the smoothed energy of the second of the two versions.

7. (Amended) Method according to claim 6, wherein the smoothing window is an exponential window defined by a forgetting factor.

8. (Amended) Method according to claim 7, comprising the step of allocating a substantially zero value to the forgetting factor when the energy of the second of the two versions is less than a value of the order of the smoothed energy of the second of the two versions.

9. (Amended) Method according to claim 8, comprising the step of allocating a first value substantially equal to 1 to the forgetting factor when the energy of the second of the two versions is greater than said value of the order of the smooth energy multiplied by a coefficient bigger than 1, and allocating a second value lying between 0 and said first value to the forgetting factor when the energy of the second of the two versions is greater than said value of the order of the smoothed energy and less than said value of the order of the smoothed energy multiplied by said coefficient.

10. (Amended) Method according to claim 1, wherein the first and second fractions correspond substantially to attenuations of 10 dB and 60 dB, respectively.

11. (Amended) Method according to claim 1, wherein the comparison of the two different versions of the speech signal is performed on respective differences between the energies of said two versions in said frequency band and a lower bound of the energy of the denoised version of the speech signal in said frequency band.

12. (Amended) Method according to claim 11, wherein one of the two different versions of the speech signal is a non-denoised version of the speech signal.

13. (Amended) Device for detecting voice activity in a speech signal, comprising signal processing means for analyzing the speech signal in at least one frequency band, wherein the processing means comprise means for comparing two different versions of the speech signal, wherein at least one of said versions is a denoised version, obtained by taking account of estimates of noise included in the signal.

Add the following claims:

16. (New) Device according to claim 13, wherein the processing means comprise means for evaluating, in said frequency band, energies of said two different versions of the speech signal, whereby inputs of the comparison means comprise said energies or a monotonic function of said energies.

17. (New) Device according to claim 13, wherein the processing means further comprises means for performing a time smoothing of the energy of one of said versions of the speech signal, and means for comparing the energy of said version and the smoothed energy.

18. (New) Device according to claim 17, wherein the processing means comprise a voice activity detection automaton having a plurality of states including a speech state and a silence state, means for controlling transitions of the voice activity detection automaton from the speech state to the silence state based on a comparison between the energy of said version and the smoothed energy, and means for controlling transitions of the voice activity detection automaton from the silence state to the speech state based on a comparison of the two different versions of the speech signal.

19. (New) Device according to claim 13, further comprising first non-linear spectral subtraction means to provide a first of the two versions of the speech signal as a denoised version which is not less, in the spectral domain, than a first fraction of a long-term estimate representative of a noise component included in the speech signal, and second non-linear spectral subtraction means to provide a second of the two versions of the speech signal as a denoised version which is not less, in the spectral domain, than a second fraction of said long-term estimate, said second fraction being smaller than said first fraction.

20. (New) Device according to claim 19, wherein the processing means further comprises means for performing a time smoothing of the energy of each of the two versions of the speech signal, by means of a smoothing window determined by comparing an energy of the second of the two versions with the smoothing energy of the second of the two versions.

21. (New) Device according to claim 20, wherein the smoothing window is an exponential window defined by a forgetting factor.

22. (New) Device according to claim 21, wherein the processing means further comprises means for allocating a substantially zero value to the forgetting factor when the energy of the second of the two versions is less than a value of the order of the smoothed energy of the second of the two versions.

23. (New) Device according to claim 22, wherein the processing means further comprises means for allocating a first value substantially equal to 1 to the forgetting factor when the energy of the second of the two versions is greater than said value of the order of the smoothed energy multiplied by a coefficient bigger than 1, and for allocating a second value lying between 0 and said first value to the forgetting factor when the energy of the second of the two versions is greater than said value of the order of the smoothed energy and less than said value of the order of the smooth energy multiplied by said coefficient.

24. (New) Device according to claim 13, wherein the first and second fractions correspond substantially to attenuations of 10 dB and 60 dB, respectively.

25. (New) Device according to claim 13, wherein the comparison of the two different versions of the speech signal is performed on respective differences between the energies of said two versions in said frequency band and a lower bound of the energy of the denoised version of the speech signal in said frequency band.

26. (New) Device according to claim 25, wherein one of the two different versions of the speech signal is a non-denoised version of the speech signal.

27. (New) A computer program product, loadable into a memory associated with a processor, and comprising portions of code for execution by the processor to detect voice activity in an input digital speech signal in at least one frequency band, whereby the voice activity is detected on the basis of an analysis comprising the step of comparing two different versions of the speech signal, wherein at least one of said versions is a denoised version obtained by taking account of estimates of noise included in the signal.

28. (New) A computer program product according to claim 27, wherein said comparison is performed on respective energies, evaluated in said frequency band, of the two different versions of the speech signal, or to a monotonic function of said energies.

29. (New) A computer program product according to claim 1, wherein said analysis further comprises a time smoothing of the energy of one of said versions of the speech signal, and a comparison between the energy of said version of the smoothed energy.

30. (New) A computer program product according to claim 29, wherein the comparison between the energy of said version and the smoothed energy control transitions of a voice activity detection automaton from a speech state to a silence state, and wherein the comparison of the two different versions of the speech signal controls transitions of the detection automaton from the silence state to the speech state.

31. (New) A computer program product according to claim 27, wherein the two different versions of the speech signal are two versions denoised by non-linear spectral subtraction, wherein a first of the two versions is denoised in such a way as not to be less, in the spectral domain, than a first fraction of a long-term estimate representative of a noise component included in the speech signal, and the second of the two versions is denoised in such a way as not to be less, in the spectral domain, than a second fraction of said long-term estimate, smaller than said first fraction.

32. (New) A computer program product according to claim 31, wherein said analysis further comprises a time smoothing of the energy of each of the two versions of the speech signal, by means of a smoothing window determined by comparing the energy of the second of the two versions with the smoothed energy of the second of the two versions.

33. (New) A computer program product according to claim 32, wherein the smoothing window is an exponential window defined by a forgetting factor.

34. (New) A computer program product according to claim 33, wherein said analysis further comprises the step of allocating a substantially zero value to the forgetting factor when the energy of the second of the two versions is less than a value of the order of the smoothed energy of the second of the two versions.

35. (New) A computer program product according to claim 34, wherein said analysis further comprises the steps of allocating a first value substantially equal to 1 to the forgetting factor when the energy of the second of the two versions is greater than said value of the order of the smoothed energy multiplied by a coefficient bigger than 1, and allocating a second value lying between 0 and said first value to the forgetting factor when the energy of the second of the two versions is greater than said value of the order of the smoothed energy and less than said value of the order of the smoothed energy multiplied by said coefficient.

36. (New) A computer program product according to claim 27, wherein the first and second fractions correspond substantially to attenuations of 10 dB and 60 dB, respectively.

37. (New) A computer program product according to claim 27, wherein the comparison of the two different versions of the speech signal is performed on respective differences between the energies of said two versions in said frequency band and a lower bound of the energy of the denoised version of the speech signal in said frequency band.

38. (New) A computer program product according to claim 37, wherein one of the two different versions of the speech signal is a non-denoised version of the speech signal.

Remarks:

Allowance of all claims is respectfully requested. The Commissioner is authorized to charge any additional fees under 37 C.F.R. § 1.16 and § 1.17, or credit any overpayment to Deposit Account No. 20-1504 (MATR-0018-US).

Date: 4/3/01

Respectfully submitted,

Dan C. Hu, Registration No. 40,025
TROP, PRUNER & HU, P.C.
8554 Katy Freeway, Suite 100
Houston, Texas 77024-1805
(713) 468-8880 [Phone]
(713) 468-8883 [Fax]

VERSIONS WITH MARKINGS TO SHOW CHANGES

IN THE CLAIMS:

Claims 14-15 have been cancelled. New claims 16-38 have been added.

Amendments of the claims are indicated below:

1. (Amended) Method for detecting voice activity in a digital speech signal [(s)] in at least one frequency band, [characterized in that] wherein the voice activity is detected on the basis of an analysis comprising [a comparison, in the said frequency band,] the step of comparing two different versions of the speech signal, [one] wherein at least one of [which] said versions is a denoised version obtained by taking account of estimates of [the] noise included in the signal.

2. (Amended) Method according to claim 1, [in which the] wherein said comparison [pertains to] is performed on respective energies [($E_{1,n,j}$, $E_{2,n,j}$)], evaluated in [the] said frequency band, of the two different versions of the speech signal, or to a monotonic function of [the] said energies.

3. (Amended) Method according to claim 1, wherein [or 2, in which the] said analysis [furthermore] further comprises a [temporal] time smoothing of the energy [($E_{1,n,j}$)] of one of [the] said versions of the speech signal, and a comparison between the energy of [the] said version and the smooth energy [($\bar{E}_{1,n,j}$)].

4. (Amended) Method according to claim 3, [in which] wherein the comparison between the energy of [the] said version [($E_{1,n,j}$)] and the smooth energy [($\bar{E}_{1,n,j}$)] controls [the] transitions of a voice activity detection automaton from a speech state to a silence state, [whilst] and wherein the comparison of the two different versions of the speech signal controls [the] transitions of the detection automaton from the silence state to the speech state.

5. (Amended) Method according to [any one of claims 1 to 4, in which] claim 1, wherein the two different versions of the speech signal are two versions denoised by non-linear spectral subtraction, wherein a first of the two versions [($\hat{E}p_{1,n,i}$) being] is denoised in such a way as not to be less, in the spectral domain, than a first fraction [($\beta 1_i$)] of a long-term estimate [($\hat{B}_{n,i}$)] representative of a noise component included in the speech signal, and the

second of the two versions [($\hat{E}p_{2,n,i}$) being] is denoised in such a way as not to be less, in the spectral domain, than a second fraction [(ß2_j)] of [the] said long-term estimate, smaller than [the] said first fraction.

6. (Amended) Method according to claim 5, [in which a temporal] wherein said analysis further comprises a time smoothing of the energy of each of the two versions of the speech signal [is performed], by means of a [determined] smoothing window determined by comparing the energy [($E_{2,n,j}$)] of the second of the two versions with the smoothed energy [$\bar{E}_{2,n,j}$)] of the second of the two versions.

7. (Amended) Method according to claim 6, [in which] wherein the smoothing window is an exponential window defined by a [forget] forgetting factor [(λ)].

8. (Amended) Method according to claim 7, [in which the forget factor (λ) has] comprising the step of allocating a substantially zero value [(λ_r)] to the forgetting factor when the energy [($E_{2,n,j}$)] of the second of the two versions is less than a value of the order of the smoothed energy [$\bar{E}_{2,n,j}$)] of the second of the two versions.

9. (Amended) Method according to claim 8, [in which the forget factor (λ) has] comprising the step of allocating a first value [(λ_q)] substantially equal to 1 to the forgetting factor when the energy [($E_{2,n,j}$)] of the second of the two versions is greater than [the] said value of the order of the smooth energy multiplied by a coefficient [(Δ)] bigger than 1, and allocating a second value [(λ_p)] lying between 0 and [the] said first value to the forgetting factor when the energy of the second of the two versions is greater than [the] said value of the order of the smoothed energy and less than [the] said value of the order of the smoothed energy multiplied by [the] said coefficient.

10. (Amended) Method according to claim 1, wherein [any one of claims 5 to 9, in which] the first and second fractions [(ß1_j, ß2_j)] correspond substantially to attenuations of 10 dB and 60 dB, respectively.

11. (Amended) Method according to claim 1, wherein [any one of claims 1 to 10, in which] the comparison of the two different versions of the speech signal [pertains to] is

performed on respective differences between the energies [($E_{1,n,j}$, $E_{2,n,j}$) of these] of said two versions in [the] said frequency band and a lower bound [($E_{2min,j}$)] of the energy [($E_{2,n,j}$)] of the denoised version of the speech signal in [the] said frequency band.

12. (Amended) Method according to claim 11, [in which] wherein one of the two different versions of the speech signal is a non-denoised version of the speech signal.

13. (Amended) Device for detecting voice activity in a speech signal, comprising signal processing means [(15) designed to implement a method according to any one of claims 1 to 12] for analyzing the speech signal in at least one frequency band, wherein the processing means comprise means for comparing two different versions of the speech signal, wherein at least one of said versions is a denoised version, obtained by taking account of estimates of noise included in the signal.

16. (New) Device according to claim 13, wherein the processing means comprise means for evaluating, in said frequency band, energies of said two different versions of the speech signal, whereby inputs of the comparison means comprise said energies or a monotonic function of said energies.

17. (New) Device according to claim 13, wherein the processing means further comprises means for performing a time smoothing of the energy of one of said versions of the speech signal, and means for comparing the energy of said version and the smoothed energy.

18. (New) Device according to claim 17, wherein the processing means comprise a voice activity detection automaton having a plurality of states including a speech state and a silence state, means for controlling transitions of the voice activity detection automaton from the speech state to the silence state based on a comparison between the energy of said version and the smoothed energy, and means for controlling transitions of the voice activity detection automaton from the silence state to the speech state based on a comparison of the two different versions of the speech signal.

19. (New) Device according to claim 13, further comprising first non-linear spectral subtraction means to provide a first of the two versions of the speech signal as a denoised version which is not less, in the spectral domain, than a first fraction of a long-term

estimate representative of a noise component included in the speech signal, and second non-linear spectral subtraction means to provide a second of the two versions of the speech signal as a denoised version which is not less, in the spectral domain, than a second fraction of said long-term estimate, said second fraction being smaller than said first fraction.

20. (New) Device according to claim 19, wherein the processing means further comprises means for performing a time smoothing of the energy of each of the two versions of the speech signal, by means of a smoothing window determined by comparing an energy of the second of the two versions with the smoothing energy of the second of the two versions.

21. (New) Device according to claim 20, wherein the smoothing window is an exponential window defined by a forgetting factor.

22. (New) Device according to claim 21, wherein the processing means further comprises means for allocating a substantially zero value to the forgetting factor when the energy of the second of the two versions is less than a value of the order of the smoothed energy of the second of the two versions.

23. (New) Device according to claim 22, wherein the processing means further comprises means for allocating a first value substantially equal to 1 to the forgetting factor when the energy of the second of the two versions is greater than said value of the order of the smoothed energy multiplied by a coefficient bigger than 1, and for allocating a second value lying between 0 and said first value to the forgetting factor when the energy of the second of the two versions is greater than said value of the order of the smoothed energy and less than said value of the order of the smooth energy multiplied by said coefficient.

24. (New) Device according to claim 13, wherein the first and second fractions correspond substantially to attenuations of 10 dB and 60 dB, respectively.

25. (New) Device according to claim 13, wherein the comparison of the two different versions of the speech signal is performed on respective differences between the energies of said two versions in said frequency band and a lower bound of the energy of the denoised version of the speech signal in said frequency band.

26. (New) Device according to claim 25, wherein one of the two different versions of the speech signal is a non-denoised version of the speech signal.

27. (New) A computer program product, loadable into a memory associated with a processor, and comprising portions of code for execution by the processor to detect voice activity in an input digital speech signal in at least one frequency band, whereby the voice activity is detected on the basis of an analysis comprising the step of comparing two different versions of the speech signal, wherein at least one of said versions is a denoised version obtained by taking account of estimates of noise included in the signal.

28. (New) A computer program product according to claim 27, wherein said comparison is performed on respective energies, evaluated in said frequency band, of the two different versions of the speech signal, or to a monotonic function of said energies.

29. (New) A computer program product according to claim 1, wherein said analysis further comprises a time smoothing of the energy of one of said versions of the speech signal, and a comparison between the energy of said version of the smoothed energy.

30. (New) A computer program product according to claim 29, wherein the comparison between the energy of said version and the smoothed energy control transitions of a voice activity detection automaton from a speech state to a silence state, and wherein the comparison of the two different versions of the speech signal controls transitions of the detection automaton from the silence state to the speech state.

31. (New) A computer program product according to claim 27, wherein the two different versions of the speech signal are two versions denoised by non-linear spectral subtraction, wherein a first of the two versions is denoised in such a way as not to be less, in the spectral domain, than a first fraction of a long-term estimate representative of a noise component included in the speech signal, and the second of the two versions is denoised in such a way as not to be less, in the spectral domain, than a second fraction of said long-term estimate, smaller than said first fraction.

32. (New) A computer program product according to claim 31, wherein said analysis further comprises a time smoothing of the energy of each of the two versions of the speech signal, by means of a smoothing window determined by comparing the energy of the second of the two versions with the smoothed energy of the second of the two versions.

33. (New) A computer program product according to claim 32, wherein the smoothing window is an exponential window defined by a forgetting factor.

34. (New) A computer program product according to claim 33, wherein said analysis further comprises the step of allocating a substantially zero value to the forgetting factor when the energy of the second of the two versions is less than a value of the order of the smoothed energy of the second of the two versions.

35. (New) A computer program product according to claim 34, wherein said analysis further comprises the steps of allocating a first value substantially equal to 1 to the forgetting factor when the energy of the second of the two versions is greater than said value of the order of the smoothed energy multiplied by a coefficient bigger than 1, and allocating a second value lying between 0 and said first value to the forgetting factor when the energy of the second of the two versions is greater than said value of the order of the smoothed energy and less than said value of the order of the smoothed energy multiplied by said coefficient.

36. (New) A computer program product according to claim 27, wherein the first and second fractions correspond substantially to attenuations of 10 dB and 60 dB, respectively.

37. (New) A computer program product according to claim 27, wherein the comparison of the two different versions of the speech signal is performed on respective differences between the energies of said two versions in said frequency band and a lower bound of the energy of the denoised version of the speech signal in said frequency band.

38. (New) A computer program product according to claim 37, wherein one of the two different versions of the speech signal is a non-denoised version of the speech signal.

09/806756

JC08 Rec'd PCT/PTO 03 APR 2001

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Application No. :

U.S. National Serial No. :

Filed :

PCT International Application No. : PCT/FR00/02220

VERIFICATION OF A TRANSLATION

I, the below named translator, hereby declare that:

My name and post office address are as stated below;

That I am knowledgeable in the French language in which the below identified international application was filed, and that, to the best of my knowledge and belief, the English translation of the international application No. PCT/FR00/02220 is a true and complete translation of the above identified international application as filed.

I hereby declare that all the statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the patent application issued thereon.

Date: March 19, 2001

DLawson

Full name of the translator : David LAWSON

For and on behalf of RWS Group plc

Post Office Address : Europa House, Marsham Way,

Gerrards Cross, Buckinghamshire,

England.

Express Mail No. EL732849804US

METHOD AND DEVICE FOR DETECTING VOICE ACTIVITY

The present invention relates to digital techniques for processing speech signals. It relates more particularly to the techniques utilizing voice activity detection so as to perform different processings depending on whether the signal does or does not carry voice activity.

The digital techniques in question come under varied domains: coding of speech for transmission or storage, speech recognition, noise reduction, echo cancellation, etc.

The main difficulty with processes for detecting voice activity is that of distinguishing between voice activity and the noise which accompanies the speech signal.

The document WO99/14737 describes a method of detecting voice activity in a digital speech signal processed on the basis of successive frames and in which an a priori denoising of the speech signal of each frame is carried out on the basis of noise estimates obtained during the processing of one or more previous frames, and the variations in the energy of the a priori denoised signal are analyzed so as to detect a degree of voice activity of the frame. By carrying out the detection of voice activity on the basis of an a priori denoised signal, the performance of this detection is substantially improved when the surrounding noise is relatively strong.

In the methods customarily used to detect voice activity, the energy variations of the (direct or denoised) signal are analyzed with respect to a long-term average of the energy of this signal, a relative increase in the instantaneous energy suggesting the appearance of voice activity.

An aim of the present invention is to propose another type of analysis allowing voice activity

detection which is robust to the noise which may accompany the speech signal.

According to the invention, there is proposed a method for detecting voice activity in a digital speech signal in at least one frequency band, whereby the voice activity is detected on the basis of an analysis comprising a comparison, in the said frequency band, of two different versions of the speech signal, one at least of which is a denoised version obtained by taking account of estimates of the noise included in the signal.

This method can be executed over the entire frequency band of the signal, or on a subband basis, as a function of the requirements of the application using voice activity detection.

Voice activity can be detected in a binary manner for each band, or measured by a continuously varying parameter which may result from the comparison between the two different versions of the speech signal.

The comparison typically pertains to respective energies, evaluated in the said frequency band, of the two different versions of the speech signal, or to a monotonic function of these energies.

Another aspect of the present invention relates to a device for detecting voice activity in a speech signal, comprising signal processing means designed to implement a method as defined hereinabove.

The invention further relates to a computer program, loadable into a memory associated with a processor, and comprising portions of code for implementing a method as defined hereinabove upon the execution of the said program by the processor, as well as to a computer medium, on which such a program is recorded.

Other features and advantages of the present invention will become apparent in the following

description of non-limiting exemplary embodiments, with reference to the appended drawings, in which:

- Figure 1 is a schematic diagram of a signal processing chain using a voice activity detector according to the invention;
- Figure 2 is a schematic diagram of an exemplary voice activity detector according to the invention;
- Figures 3 and 4 are flow charts of signal processing operations performed in the detector of Figure 2;
- Figure 5 is a graphic showing an exemplary profile of energies calculated in the detector of Figure 2 and illustrating the principle of voice activity detection;
- Figure 6 is a diagram of a detection automaton implemented in the detector of Figure 2;
- Figure 7 is a schematic diagram of another embodiment of a voice activity detector according to the invention;
- Figure 8 is a flow chart of signal processing operations performed in the detector of Figure 7;
- Figure 9 is a graphic of a function used in the operations of Figure 8.

The device of Figure 1 processes a digital speech signal s. The signal processing chain represented produces voice activity decisions $\delta_{n,j}$ which are usable in a manner known per se by application units, not represented, affording functions such as speech coding, speech recognition, noise reduction, echo cancellation, etc. The decisions $\delta_{n,j}$ can comprise a frequency resolution (index j), this making it possible to enhance applications operating in the frequency domain.

A windowing module 10 puts the signal s into the form of successive windows or frames of index n, each consisting of a number N of samples of digital signal. In a conventional manner, these frames may

exhibit mutual overlaps. In the remainder of the present description, the frames will be regarded, without this being in any way limiting, as consisting of N = 256 samples at a sampling frequency F_e of 8 kHz, with a Hamming weighting in each window, and overlaps of 50% between consecutive windows.

The signal frame is transformed into the frequency domain by a module 11 applying a conventional fast Fourier transform algorithm (FFT) for calculating the modulus of the spectrum of the signal. The module 11 then delivers a set of N = 256 frequency components of the speech signal, which are denoted $S_{n,f}$, where n designates the current frame number, and f a frequency of the discrete spectrum. Owing to the properties of digital signals in the frequency domain, only the first N/2 = 128 samples are used.

To calculate the estimates of the noise contained in the signal s, we do not use the frequency resolution available at the output of the fast Fourier transform, but a lower resolution, determined by a number I of frequency subbands covering the $[0, F_e/2]$ band of the signal. Each subband i ($1 \leq i \leq I$) extends between a lower frequency f(i-1) and an upper frequency f(i), with f(0) = 0, and $f(I) = F_e/2$. This chopping into subbands can be uniform ($f(i)-f(i-1) = F_e/2I$). It may also be non-uniform (for example according to a barks scale). A module 12 calculates the respective averages of the spectral components $S_{n,f}$ of the speech signal on a subband basis, for example through a uniform weighting such as:

$$S_{n,i} = \frac{1}{f(i)-f(i-1)} \sum_{f \in [f(i-1),f(i)[} S_{n,f}$$

This averaging reduces the fluctuations between the subbands by averaging the contributions of the noise in these subbands, and this will reduce the

variance of the noise estimator. Furthermore, this averaging makes it possible to reduce the complexity of the system.

The averaged spectral components $S_{n,i}$ are addressed to a voice activity detection module 15 and to a noise estimation module 16. $\hat{B}_{n,i}$ denotes the long-term estimate of the noise component produced by the module 16 in relation to frame n and to subband i.

These long-term estimates $\hat{B}_{n,i}$ may for example be obtained in the manner described in WO99/14737. It is also possible to use simple smoothing by means of an exponential window defined by a forget factor λ_B:

$$\hat{B}_{n,i} = \lambda_B.\hat{B}_{n-1,i} + (1-\lambda_B).S_{n,i}$$

with λ_B equal to 1 if the voice activity detector 15 indicates that subband i bears voice activity, and equal to a value lying between 0 and 1 otherwise.

Of course, it is possible to use other long-term estimates representative of the noise component included in the speech signal, these estimates may represent a long-term average, or else a minimum of the component $S_{n,j}$ over a sufficiently long sliding window.

Figures 2 to 6 illustrate a first embodiment of the voice activity detector 15. A denoising module 18 executes, for each frame n and each subband i, the operations corresponding to steps 180 to 187 of Figure 3, so as to produce two denoised versions $\hat{E}p_{1,n,i}$, $\hat{E}p_{2,n,i}$ of the speech signal. This denoising is done by non-linear spectral subtraction. The first version $\hat{E}p_{1,n,i}$, is denoised in such a way as not to be less, in the spectral domain, than a fraction $ß1_i$ of the long-term estimate $\hat{B}_{n-\tau 1,i}$. The second version $\hat{E}p_{2,n,i}$ is denoised in such a way as not to be less, in the spectral domain, than a fraction $ß2_j$ of the long-term estimate $\hat{B}_{n-\tau 1,i}$. The quantity $\tau 1$ is a delay expressed as a number of frames, which may be fixed (for example $\tau 1 = 1$) or variable. The more confident one is in the voice activity detection, the smaller the delay will be. The

fractions $\beta1_i$ and $\beta2_i$ (such that $\beta1_i > \beta2_i$) may be dependent on or independent of subband i. Preferred values correspond for $\beta1_i$ to an attenuation of 10 dB, and for $\beta2_i$ to an attenuation of 60 dB, i.e. $\beta1_i \approx 0.3$ and $\beta2_i \approx 0.001$.

In step 180, the module 18 calculates, with the resolution of the subbands i, the frequency response $Hp_{n,i}$ of the a priori denoising filter, according to:

$$Hp_{n,i} = \frac{S_{n,i} - \alpha'_{n-\tau1,i}.\hat{B}_{n-\tau1,i}}{S_{n-\tau2,i}}$$

where $\tau2$ is a positive or zero integer delay and $\alpha'_{n,i}$ is a noise overestimation coefficient. This overestimation coefficient $\alpha'_{n,i}$ may be dependent on or independent of the frame index n and/or the subband index i. In a preferred embodiment, it depends both on n and i, and it is determined as described in document WO99/14737. A first denoising is performed in step 181: $\hat{E}p_{n,i} = Hp_{n,i}.S_{n,i}$. In steps 182 to 184, the spectral components $\hat{E}p_{1,n,i}$ are calculated according $\hat{E}p_{1,n,i} = \max(\hat{E}p_{n,i} : \beta1_i.\hat{B}_{n-\tau1,i})$, and in steps 182 to 184, the spectral components $\hat{E}p_{2,n,i}$ are calculated according to $\hat{E}p_{2,n,i} = \max(\hat{E}p_{n,i} : \beta2_i.\hat{B}_{n-\tau1,i})$

The voice activity detector 15 of Figure 2 comprises a module 19 which calculates energies of the denoised versions of the signal $\hat{E}p_{1,n,i}$ and $\hat{E}p_{2,n,i}$ respectively lying in m frequency bands designated by the index $j (1 \leq j \leq m, m \geq 1)$. This resolution may be the same as that of the subbands defined by the module 12 (index i), or a finer resolution of possibly as much as the whole of the useful band $[0, F_e/2]$ of the signal (case m = 1). By way of example, the module 12 can define I = 16 uniform subbands of the band $[0, F_e/2]$, and the module 19 can retain m = 3 wider bands, each band of index j covering the subbands of index i ranging from imin(j) to imax(j), with imin(1) = 1, imin(j+1) = imax(j) + 1 for $1 \leq j < m$, and

imax(m) = 1. In step 190 (Figure 3), the module 19 calculates the energies per band:

$$E_{1,n,j} = \sum_{i=imin(j)}^{imax(j)} [f(i) - f(i-1)].\hat{E}p_{1,n,i}^{2}$$

$$E_{2,n,j} = \sum_{i=imin(j)}^{imax(j)} [f(i) - f(i-1)].\hat{E}p_{2,n,i}^{2}$$

A module 20 of the voice activity detector 15 performs a temporal smoothing of the energies $E_{1,n,j}$ and $E_{2,n,j}$ for each of the bands of index j, this corresponding to steps 200 to 205 for Figure 4. The smoothing of these two energies is performed by means of a determined smoothing window by comparing the energy $E_{2,n,j}$ of the most denoised version with its previously calculated smoothed energy $\overline{E}_{2,n-1,j}$, or with a value of the order of this smoothed energy $\overline{E}_{2,n-1,j}$, (tests 200 and 201). This smoothing window can be an exponential window defined by a forget factor λ lying between 0 and 1. This forget factor λ can take three values: the one λ_r very close to 0 (for example $\lambda_r = 0$) chosen in step 202 if $E_{2,n,j} \leq \overline{E}_{2,n-1,j}$; the second λ_q very close to 1 (for example $\lambda_q = 0.99999$) chosen in step 203 if $E_{2,n,j} > \Delta\ \overline{E}_{2,n-1,j}$, Δ being a coefficient bigger than 1; and the third λ_p lying between 0 and λ_q (for example $\lambda_p = 0.98$) chosen in step 204 if $\overline{E}_{2,n-1,j} < E_{2,n-1,j} \leq \Delta\ \overline{E}_{2,n-1,j}$. The exponential smoothing with the forget factor λ is then performed conventionally in step 205 according to:

$$\overline{E}_{1,n,j} = \lambda.\overline{E}_{1,n-1,j} + (1-\lambda).E_{1,n,j}$$

$$\overline{E}_{2,n,j} = \lambda.\overline{E}_{2,n-1,j} + (1-\lambda).E_{2,n,j}$$

An exemplary variation over time of the energies $E_{1,n,j}$ and $E_{2,n,j}$ and of the smoothed energies $\overline{E}_{1,n,j}$ and $\overline{E}_{2,n,j}$ is represented in Figure 5. It may be seen that good tracking of the smoothed energies is

achieved when the forget factor is determined on the basis of the variations in the energy $E_{2,n,j}$ corresponding to the most denoised version of the signal. The forget factor λ_p makes it possible to take into account the increases in the level of the background noise, the energy reductions being tracked by the forget factor λ_r. The forget factor λ_q very close to 1 means that the smoothed energies do not track the abrupt energy increases due to speech. However, the factor λ_q remains slightly less than 1 so as to avoid errors caused by an increase in the background noise which may arise during a fairly long period of speech.

The voice activity detection automaton is controlled in particular by a parameter resulting from a comparison of the energies $E_{1,n,j}$ and $E_{2,n,j}$. This parameter can in particular be the ratio $d_{n,j} = E_{1,n,j}/E_{2,n,j}$. It may be seen in Figure 5 that this ratio $d_{n,j}$ allows proper detection of the speech phases (represented by hatching).

The control of the detection automaton can also use other parameters, such as a parameter related to the signal-to-noise ratio: $snr_{n,j} = E_{1,n,j}/\overline{E}_{1,n,j}$, this amounting to taking into account a comparison between the energies $E_{1,n,j}$ and $\overline{E}_{1,n,j}$. The module 21 for controlling the automata relating to the various bands of index j calculates the parameters $d_{n,j}$ and $snr_{n,j}$ in step 210, then determines the state of the automata. The new state $\delta_{n,j}$ of the automaton relating to band j depends on the previous state $\delta_{n-1,j}$, on $d_{n,j}$ and on $snr_{n,j}$, for example as indicated in the diagram of Figure 6.

Four states are possible: $\delta_j = 0$ detects silence, or absence of speech; $\delta_j = 2$ detects the presence of voice activity; and the states $\delta_j = 1$ and $\delta_j = 3$ are intermediate states of ascent and descent. When the automaton is in the silence state ($\delta_{n-1,j} = 0$), it remains there if $d_{n,j}$ exceeds a first threshold $\alpha 1_j$, and if it switches to the ascent state in the converse

case. In the ascent state ($\delta_{n-1,j} = 1$), it returns to the silence state if $d_{n,j}$ exceeds a second threshold $\alpha 2_j$; and it switches to the speech state in the converse case. When the automaton is in the speech state ($\delta_{n-1,j} = 2$), it remains there if $snr_{n,j}$ exceeds a third threshold $\alpha 3_j$, and it switches to the descent state in the converse case. In the descent state ($\delta_{n-1,j} = 3$), the automaton returns to the speech state if $snr_{n,j}$ exceeds a fourth threshold $\alpha 4_j$, and it returns to the silence state in the converse case. The thresholds $\alpha 1_j$, $\alpha 2_j$, $\alpha 3_j$, and $\alpha 4_j$ may be optimized separately for each of the frequency bands j.

It is also possible for the automata relating to the various bands to be made to interact by the module 21.

In particular, it may force each of the automata relating to each of the subbands to the speech state as soon as one among them is in the speech state. In this case, the output of the voice activity detector 15 relates to the whole of the signal band.

The two appendices to the present description show a source code in the C++ language, with a fixed-point data representation corresponding to an implementation of the exemplary voice activity detection method described hereinabove. To embody the detector, one possibility is to translate this source code into executable code, to record it in a program memory associated with an appropriate signal processor, and to have it executed by this processor on the input signals of the detector. The function *a_priori_signal_power* presented in appendix 1 corresponds to the operations incumbent on the modules 18 and 19 of the voice activity detector 15 of Figure 2. The function *voice_activity_detector* presented in appendix 2 corresponds to the operations incumbent on the modules 20 and 21 of this detector.

In the particular example of the appendices, the following parameters have been employed: $\tau 1 = 1$;

$\tau2 = 0$; $ß1_i = 0.3$; $ß2_i = 0.001$; $m = 3$; $\Delta = 4.953$; $\lambda_p = 0.98$; $\lambda_q = 0.99999$; $\lambda_r = 0$; $\alpha1_j = \alpha2_j = \alpha4_j = 1.221$; $\alpha3_j = 1.649$. Table 1 hereinbelow gives the correspondences between the notation employed in the above description and in the drawings and that employed in the appendix.

subband	I
E[subband]	$S_{n,i}$
module	$\hat{E}p_{n,i}$ or $\hat{E}p_{1,n,i}$ or $\hat{E}p_{2,n,i}$
param.beta_a_priori1	$ß1_j$
param.beta_a_priori2	$ß2_j$
vad	j-1
param.vad_number	m
P1[vad]	$E_{1,n,j-1}$
P1s[vad]	$\overline{E}_{1,n-j,1}$
P2[vad]	$E_{2,n,j-1}$
P2s[vad]	$\overline{E}_{2,n,j-1}$
DELTA_P	$\mathrm{Log}(\Delta)$
d	$\mathrm{Log}(d_{n,j})$
snr	$\mathrm{Log}(snr_{n,j})$
NOISE	silence state
ASCENT	ascent state
SIGNAL	speech state
DESCENT	descent state
D_NOISE	$\mathrm{Log}(\alpha1_j)$
D_SIGNAL	$\mathrm{Log}(\alpha2_j)$
SNR_SIGNAL	$\mathrm{LOG}(\alpha3_j)$
SNR_NOISE	$\mathrm{Log}(\alpha4_j)$

TABLE I

In the variant embodiment illustrated by Figure 7, the denoising module 25 of the voice activity detector 15 delivers a single denoised version $\hat{E}p_{n,i}$ of the speech signal, so that the module 26 calculates its energy $E_{2,n,j}$ for each band j. The other version, in

which the module 26 calculates the energy, is represented directly by the non-denoised samples $S_{n,i}$.

As before, various denoising processes may be applied by the module 25. In the example illustrated by steps 250 to 256 of Figure 8, the denoising is done by nonlinear spectral subtraction with a noise overestimation coefficient dependent on a quantity ρ related to the signal-to-noise ratio. In steps 250 to 252, a preliminary denoising is performed for each subband of index i according to :

$$S'_{n,i} = \max\left(S_{n,i} - \alpha.\hat{B}_{n-1,i} ; \beta.\hat{B}_{n-1,i}\right),$$

the preliminary overestimation coefficient being for example $\alpha = 2$, and the fraction ß possibly corresponding to a noise attenuation of the order of 10 dB.

The quantity ρ is taken equal to the ratio $S'_{n,i}/S_{n,i}$ in step 253. The overestimation factor $f(\rho)$ varies in a nonlinear manner with the quantity ρ, for example as represented in Figure 9. For the values of ρ closest to 0 ($\rho < \rho_1$), the signal-to-noise ratio is low, and it is possible to take an overestimation factor $f(\rho) = 2$. For the highest values of ρ ($\rho_2 \leq \rho \leq 1$), the noise is weak and need not be overestimated ($f(\rho) = 1$). Between ρ_1 and ρ_2, $f(\rho)$ decreases from 2 to 1, for example linearly. The denoising proper, providing the version $\hat{E}p_{n,i}$, is performed in steps 254 to 256:

$$\hat{E}p_{n,i} = \max\left(S_{n,i} - f(\rho).\hat{B}_{n-1,i} ; \beta.\hat{B}_{n-1,i}\right).$$

The voice activity detector 15 considered with reference to Figure 7 uses, in each frequency band of index j (and/or in full band), a detection automaton having two states, silence or speech. The energies $E_{1,n,j}$ and $E_{2,n,j}$ calculated by the module 26 are respectively those contained in the components $S_{n,i}$ of the speech signal and those contained in the denoised components $\hat{E}p_{n,i}$ calculated over the various bands as indicated in step 260 of Figure 8. The comparison of the two

different versions of the speech signal pertains to respective differences between the energies $E_{1,n,j}$ and $E_{2,n,j}$ and a lower bound of the energy $E_{2,n,j}$ of the denoised version.

This lower bound $E_{2min,j}$ can in particular correspond to a minimum value, over a sliding window, of the energy $E_{2,n,j}$ of the denoised version of the speech signal in the frequency band considered. In this case, a module 27 stores in a memory of the first-in first-out type (FIFO) the L most recent values of the energy $E_{2,n,j}$ of the denoised signal in each band j, over a sliding window representing for example of the order of 20 frames, and delivers the minimum energies $E_{2min,j} = \min_{0 \le k < L} E_{2,n-k,j}$ over this window (step 270 of Figure 8). In each band, this minimum energy $E_{2min,j}$ serves as lower bound for the module 28 for controlling the detection automaton, which uses a measure M_j given by $M_j = \dfrac{E_{2,n,j} - E_{2min,j}}{E_{1,n,j} - E_{2min,j}}$ (step 280).

The automaton can be a simple binary automaton using a threshold A_j, possibly dependent on the band considered: If $M_j > A_j$, the output bit $\delta_{n,j}$ of the detector represents a silence state of the band j, and if $M_j \le A_j$, it represents a speech state. As a variant, the module 28 could deliver a nonbinary measure of the voice activity, represented by a decreasing function of M_j.

As a variant, the lower bound $E_{2min,j}$ used in step 280 could be calculated with the aid of an exponential window, with a forget factor. It could also be represented by the energy over band j of the quantity $\beta.\hat{B}_{n-1,i}$ serving as floor in the denoising by spectral subtraction.

In the foregoing, the analysis performed in order to decide on the presence or absence of voice activity pertains directly to energies of different

versions of the speech signal. Of course, the comparisons could pertain to a monotonic function of these energies, for example a logarithm, or to a quantity having similar behavior to the energies according to voice activity (for example the power).

APPENDIX 1

```
/*********************************************************************
******
 * description
 * -----------
 * NSS module:
 *  signal power before VAD
 *
 *********************************************************************
******/

/*-------------------------------------------------------------------
-----*
 *                          included files
 *-------------------------------------------------------------------
-----*/
#include <assert.h>

#include "private.h"


/*-------------------------------------------------------------------
-----*
 *                            private
 *-------------------------------------------------------------------
-----*/
Word32 power(Word16 module, Word16 beta, Word16 thd, Word16 val);


/*-------------------------------------------------------------------
-----*
 *                      a_priori_signal_power
 *-------------------------------------------------------------------
-----*/
void a_priori_signal_power
(
/* IN */      Word16 *E, Word16 *internal_state, Word16 *max_noise, W
ord16 *long_term_noise,
              Word16 *frequential_scale,

/* IN&OUT */ Word16 *alpha,

/* OUT */     Word32 *P1, Word32 *P2
)
{
   int vad;

   for(vad = 0; vad < param.vad_number; vad++) {
      int start = param.vads[vad].first_subband_for_power;
      int stop = param.vads[vad].last_subband;
      int subband;
      int uniform_subband;

      uniform_subband = 1;
```

```
      for(subband = start; subband <= stop; subband++)
         if(param.subband_size[subband] != param.subband_size[start]
)
            uniform_subband = 0;

      P1[vad] = 0; move32();
      P2[vad] = 0; move32();
      test(); if(sub(internal_state[vad], NOISE) == 0) {
         for(subband = start; subband <= stop; subband++) {
            Word32 pwr;
            Word16 shift;
            Word16 module;
            Word16 alpha_long_term;

            alpha_long_term = shr(max_noise[subband], 2); move16();
            test(); test(); if(sub(alpha_long_term, long_term_noise[
subband]) >= 0) {
               alpha[subband] = 0x7fff; move16();
               alpha_long_term = long_term_noise[subband]; move16();
            } else if(sub(max_noise[subband], long_term_noise[subban
d]) < 0) {
               alpha[subband] = 0x2000; move16();
               alpha_long_term = shr(long_term_noise[subband],2); mo
ve16();
            } else {
               alpha[subband] = div_s(alpha_long_term, long_term_noi
se[subband]); move16();
            }
            module = sub(E[subband], shl(alpha_long_term, 2)); movel
6();

            if(uniform_subband) {
               shift = shl(frequential_scale[subband], 1); move16();
            } else {
               shift = add(param.subband_shift[subband], shl(frequen
tial_scale[subband], 1)); move16();
            }

            pwr = power(module, param.beta_a_priori1, long_term_nois
e[subband], long_term_noise[subband]);
            pwr = L_shr(pwr, shift);
            P1[vad] = L_add(P1[vad], pwr); move32();

            pwr = power(module, param.beta_a_priori2, long_term_nois
e[subband], long_term_noise[subband]);
            pwr = L_shr(pwr, shift);
            P2[vad] = L_add(P2[vad], pwr); move32();
         }
      } else {
         for(subband = start; subband <= stop; subband++) {
            Word32 pwr;
            Word16 shift;
            Word16 module;
            Word16 alpha_long_term;

            alpha_long_term = mult(alpha[subband], long_term_noise[s
```

```
ubband]); move16();
              module = sub(E[subband], shl(alpha_long_term, 2)); movel
6();

              if(uniform_subband) {
                 shift = shl(frequential_scale[subband], 1); move16();
              } else {
                 shift = add(param.subband_shift[subband], shl(frequen
tial_scale[subband], 1)); move16();
              }

              pwr = power(module, param.beta_a_priori1, long_term_nois
e[subband], E[subband]);
              pwr = L_shr(pwr, shift);
              P1[vad] = L_add(P1[vad], pwr); move32();

              pwr = power(module, param.beta_a_priori2, long_term_nois
e[subband], E[subband]);
              pwr = L_shr(pwr, shift);
              P2[vad] = L_add(P2[vad], pwr); move32();
          }
       }
    }
}


/*-------------------------------------------------------------------
-----*
 *                               power
 *-------------------------------------------------------------------
-----*/
Word32 power(Word16 module, Word16 beta, Word16 thd, Word16 val)
{
   Word32 power;

   test(); if(sub(module, mult(beta, thd)) <= 0) {
      Word16 hi, lo;

      power = L_mult(val, val); move32();

      L_Extract(power, &hi, &lo);
      power = Mpy_32_16(hi, lo, beta); move32();

      L_Extract(power, &hi, &lo);
      power = Mpy_32_16(hi, lo, beta); move32();
   } else {
      power = L_mult(module, module); move32();
   }
   return(power);
}
```

APPENDIX 2

```
/*****************************************************************
******
 * description
 * -----------
 * NSS module:
 *  VAD
 *
 *****************************************************************
******/

/*----------------------------------------------------------------
-----*
 *                          included files
 *----------------------------------------------------------------
-----*/
#include <assert.h>

#include "private.h"

#include "simutool.h"


/*----------------------------------------------------------------
-----*
 *                          private
 *----------------------------------------------------------------
-----*/
#define DELTA_P         (1.6 * 1024)
#define D_NOISE         (.2 * 1024)
#define D_SIGNAL        (.2 * 1024)
#define SNR_SIGNAL      (.5 * 1024)
#define SNR_NOISE       (.2 * 1024)


/*----------------------------------------------------------------
-----*
 *                     voice_activity_detector
 *----------------------------------------------------------------
-----*/
void voice_activity_detector
(
/* IN */      Word32 *P1, Word32 *P2, Word16 frame_counter,

/* IN&OUT */ Word32 *P1s, Word32 *P2s, Word16 *internal_state,

/* OUT */    Word16 *state
)
{
   int vad;
   int signal;
   int noise;
```

```
    signal = 0; move16();
    noise = 1; move16();
    for(vad = 0; vad < param.vad_number; vad++) {
       Word16 snr, d;
       Word16 logP1, logP1s;
       Word16 logP2, logP2s;


       logP2  = logfix(P2[vad]); move16();
       logP2s = logfix(P2s[vad]); move16();

       test(); if(L_sub(P2[vad], P2s[vad]) > 0) {
          Word16 hi1, lo1;
          Word16 hi2, lo2;

          L_Extract(L_sub(P1[vad], P1s[vad]), &hi1, &lo1);
          L_Extract(L_sub(P2[vad], P2s[vad]), &hi2, &lo2);

          test(); if(sub(sub(logP2, logP2s), DELTA_P) < 0) {
             P1s[vad] = L_add(P1s[vad], L_shr(Mpy_32_16(hi1, lo1, 0x6
666), 4)); move32();
             P2s[vad] = L_add(P2s[vad], L_shr(Mpy_32_16(hi2, lo2, 0x6
666), 4)); move32();
          } else {
             P1s[vad] = L_add(P1s[vad], L_shr(Mpy_32_16(hi1, lo1, 0x6
8db), 13)); move32();
             P2s[vad] = L_add(P2s[vad], L_shr(Mpy_32_16(hi2, lo2, 0x6
8db), 13)); move32();
          }
       } else {
          P1s[vad] = P1[vad]; move32();
          P2s[vad] = P2[vad]; move32();
       }


       logP1  = logfix(P1[vad]); move16();
       logP1s = logfix(P1s[vad]); move16();


       d = sub(logP1, logP2); move16();
       snr = sub(logP1, logP1s); move16();

       ProbeFix16("d", &d, 1, 1.);
       ProbeFix16("_snr", &snr, 1, 1.);
{
   Word16 pp;
   ProbeFix16("p1", &logP1, 1, 1.);
   ProbeFix16("p2", &logP2, 1, 1.);
   ProbeFix16("p1s", &logP1s, 1, 1.);
   ProbeFix16("p2s", &logP2s, 1, 1.);
   pp = logP2 - logP2s;
   ProbeFix16("dp", &pp, 1, 1.);
}
```

```
    test(); if(sub(internal_state[vad], NOISE) == 0)
       goto LABEL_NOISE;
    test(); if(sub(internal_state[vad], ASCENT) == 0)
       goto LABEL_ASCENT;
    test(); if(sub(internal_state[vad], SIGNAL) == 0)
       goto LABEL_SIGNAL;
    test(); if(sub(internal_state[vad], DESCENT) == 0)
       goto LABEL_DESCENT;

    LABEL_NOISE:
    test(); if(sub(d, D_NOISE) < 0) {
       internal_state[vad] = ASCENT; move16();
    }
    goto LABEL_END_VAD;

    LABEL_ASCENT:
    test(); if(sub(d, D_SIGNAL) < 0) {
       internal_state(vad] = SIGNAL; move16();
       signal = 1; move16();
       noise = 0; move16();
    } else {
       internal_state[vad] = NOISE; move16();
    }
    goto LABEL_END_VAD;

    LABEL_SIGNAL:
    test(); if(sub(snr, SNR_SIGNAL) < 0) {
       internal_state[vad] = DESCENT; move16();
    } else {
       signal = 1; move16();
    }
    noise = 0; move16();
    goto LABEL_END_VAD;

    LABEL_DESCENT:
    test(); if(sub(snr, SNR_NOISE) < 0) {
       internal_state[vad] = NOISE; move16();
    } else {
       internal_state[vad] = SIGNAL; move16();
       signal = 1; move16();
       noise = 0; move16();
    }
    goto LABEL_END_VAD;

    LABEL_END_VAD:
    ;
}

*state = TRANSITION; move16();
test(); test(); if(signal != 0) {
   test(); if(sub(frame_counter, param.init_frame_number) >= 0) {
      for(vad = 0; vad < param.vad_number; vad++) {
         internal_state[vad] = SIGNAL; move16();
      }
      *state = SIGNAL; move16();
   }
```

```
        } else if(noise != 0) {
          *state = NOISE; move16();
        }
    }
```

CLAIMS

1. Method for detecting voice activity in a digital speech signal (s) in at least one frequency band, characterized in that the voice activity is detected on the basis of an analysis comprising a comparison, in the said frequency band, of two different versions of the speech signal, one at least of which is a denoised version obtained by taking account of estimates of the noise included in the signal.

2. Method according to claim 1, in which the said comparison pertains to respective energies ($E_{1,n,j}, E_{2,n,j}$), evaluated in the said frequency band, of the two different versions of the speech signal, or to a monotonic function of the said energies.

3. Method according to claim 1 or 2, in which the said analysis furthermore comprises a temporal smoothing of the energy ($E_{1,n,j}$) of one of the said versions of the speech signal, and a comparison between the energy of the said version and the smoothed energy ($\overline{E}_{1,n,j}$).

4. Method according to claim 3, in which the comparison between the energy of the said version ($E_{1,n,j}$) and the smoothed energy ($\overline{E}_{1,n,j}$) controls the transitions of a voice activity detection automaton from a speech state to a silence state, whilst the comparison of the two different versions of the speech signal controls the transitions of the detection automaton from the silence state to the speech state.

5. Method according to any one of claims 1 to 4, in which the two different versions of the speech signal are two versions denoised by non-linear spectral subtraction, a first of the two versions ($\hat{E}p_{1,n,i}$) being denoised in such a way as not to be less, in the spectral domain, than a first fraction ($ß1_i$) of a long-term estimate ($\hat{B}_{n,i}$) representative of a noise component included in the speech signal, and the second of the

two versions ($\hat{E}p_{2,n,i}$) being denoised in such a way as not to be less, in the spectral domain, than a second fraction ($ß2_j$) of the said long-term estimate, smaller than the first fraction.

6. Method according to claim 5, in which a temporal smoothing of the energy of each of the two versions of the speech signal is performed, by means of a determined smoothing window by comparing the energy ($E_{2,n,j}$) of the second of the two versions with the smoothed energy ($\bar{E}_{2,n,j}$) of the second of the two versions.

7. Method according to claim 6, in which the smoothing window is an exponential window defined by a forget factor (λ).

8. Method according to claim 7, in which the forget factor (λ) has a substantially zero value (λ_r) when the energy ($E_{2,n,j}$) of the second of the two versions is less than a value of the order of the smoothed energy ($\bar{E}_{2,n,j}$) of the second of the two versions.

9. Method according to claim 8, in which the forget factor (λ) has a first value (λ_q) substantially equal to 1 when the energy ($E_{2,n,j}$) of the second of the two versions is greater than the said value of the order of the smoothed energy multiplied by a coefficient (Δ) bigger than 1, and a second value (λ_p) lying between 0 and the said first value when the energy of the second of the two versions is greater than the said value of the order of the smoothed energy and less than the said value of the order of the smoothed energy multiplied by the said coefficient.

10. Method according to any one of claims 5 to 9, in which the first and second fractions ($ß1_j$, $ß2_j$) correspond substantially to attenuations of 10 dB and 60 dB, respectively.

11. Method according to any one of claims 1 to 10, in which the comparison of the two different versions of the speech signal pertains to respective differences

between the energies ($E_{1,n,j}$, $E_{2,n,j}$) of these two versions in the said frequency band and a lower bound ($E_{2min,j}$) of the energy ($E_{2,n,j}$) of the denoised version of the speech signal in the said frequency band.

12. Method according to claim 11, in which one of the two different versions of the speech signal is a non-denoised version of the speech signal.

13. Device for detecting voice activity in a speech signal, comprising signal processing means (15) designed to implement a method according to any one of claims 1 to 12.

14. Computer program, loadable into a memory associated with a processor, and comprising portions of code for implementing a method according to any one of claims 1 to 12 upon the execution of the said program by the processor.

15. Computer medium, on which a program according to claim 14 is recorded.

09/806756

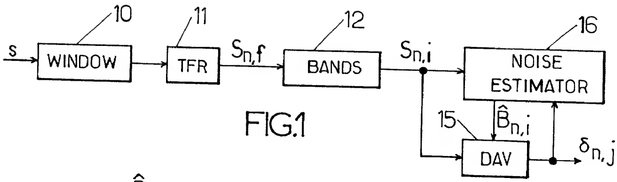


FIG.1

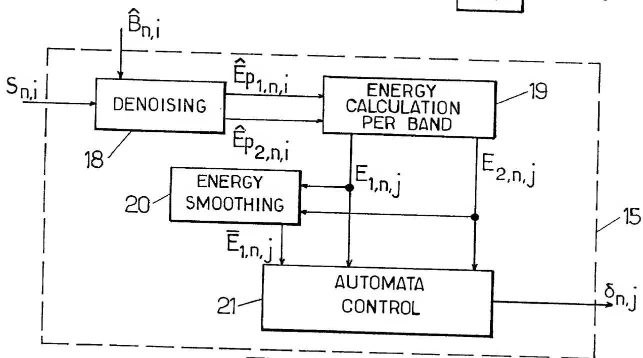
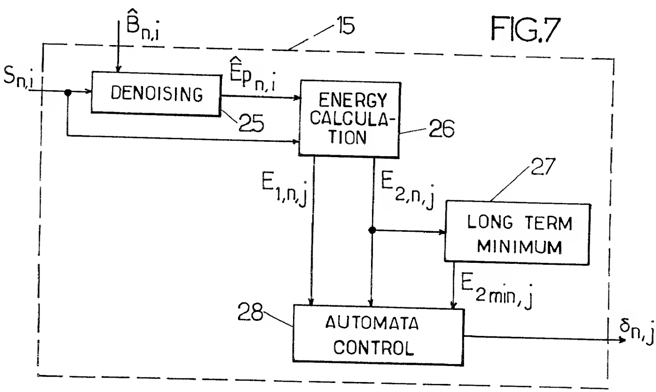


FIG.2

FIG.7



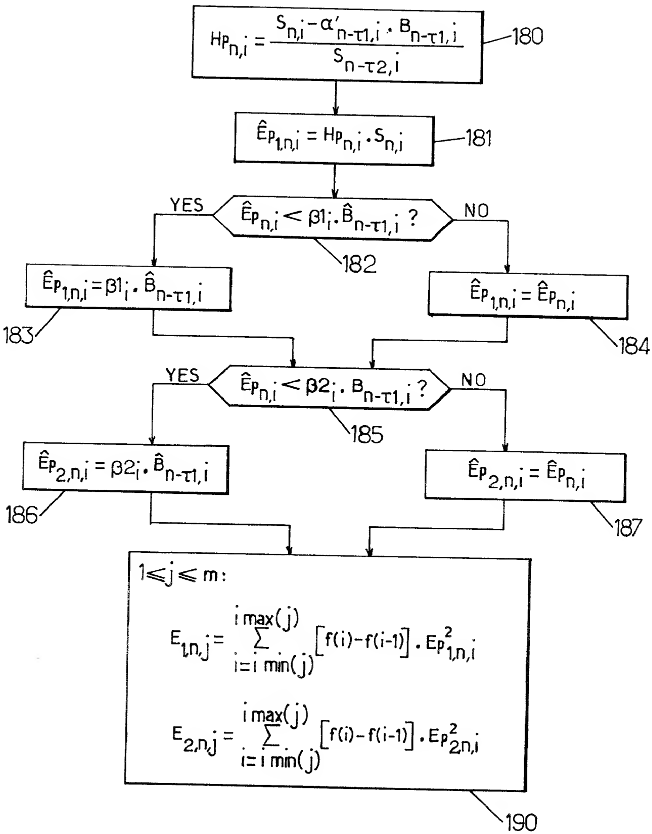
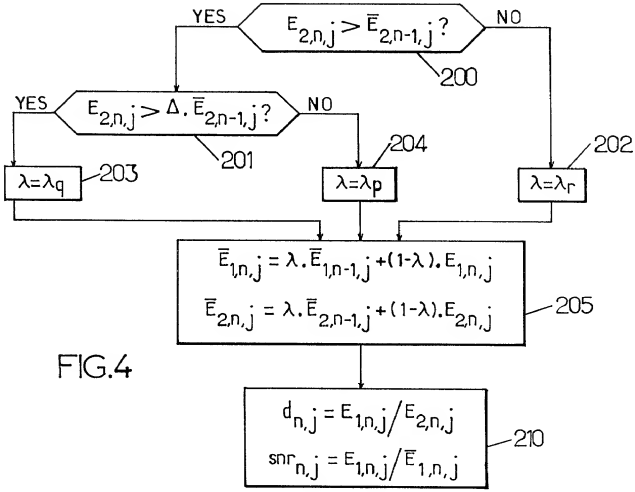

$Hp_{n,i} = \frac{S_{n,i} - \alpha'_{n-\tau1,i} \cdot B_{n-\tau1,i}}{S_{n-\tau2,i}}$
180
$\hat{E}p_{1,n,i} = Hp_{n,i} \cdot S_{n,i}$
181
YES
$\hat{E}p_{n,i} < \beta1_i \cdot \hat{B}_{n-\tau1,i}$?
NO
182
$\hat{E}p_{1,n,i} = \beta1_i \cdot \hat{B}_{n-\tau1,i}$
183
$\hat{E}p_{1,n,i} = \hat{E}p_{n,i}$
184
YES
$\hat{E}p_{n,i} < \beta2_i \cdot B_{n-\tau1,i}$?
NO
185
$\hat{E}p_{2,n,i} = \beta2_i \cdot \hat{B}_{n-\tau1,i}$
186
$\hat{E}p_{2,n,i} = \hat{E}p_{n,i}$
187
$1 \leqslant j \leqslant m$:
$E_{1,n,j} = \sum_{i=i\,min(j)}^{i\,max(j)} [f(i)-f(i-1)] \cdot Ep^2_{1,n,i}$
$E_{2,n,j} = \sum_{i=i\,min(j)}^{i\,max(j)} [f(i)-f(i-1)] \cdot Ep^2_{2,n,i}$
190

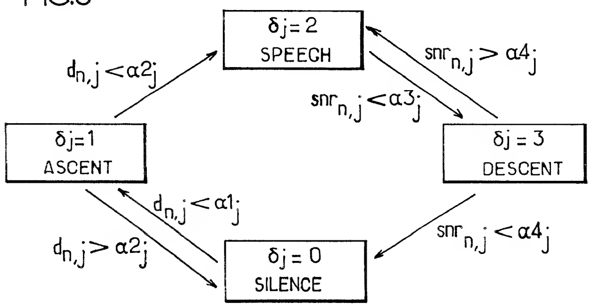
FIG.3



$E_{2,n,j} > \bar{E}_{2,n-1,j}$?
200
YES
NO
$E_{2,n,j} > \Delta . \bar{E}_{2,n-1,j}$?
201
YES
NO
$\lambda=\lambda_q$
203
$\lambda=\lambda_p$
204
$\lambda=\lambda_r$
202
$\bar{E}_{1,n,j} = \lambda . \bar{E}_{1,n-1,j} + (1-\lambda) . E_{1,n,j}$
$\bar{E}_{2,n,j} = \lambda . \bar{E}_{2,n-1,j} + (1-\lambda) . E_{2,n,j}$
205
$d_{n,j} = E_{1,n,j} / E_{2,n,j}$
$snr_{n,j} = E_{1,n,j} / \bar{E}_{1,n,j}$
210

FIG.4

FIG.6



$\delta_j = 2$
SPEECH
$\delta_j = 1$
ASCENT
$\delta_j = 3$
DESCENT
$\delta_j = 0$
SILENCE
$d_{n,j} < \alpha 2_j$
$snr_{n,j} > \alpha 4_j$
$snr_{n,j} < \alpha 3_j$
$d_{n,j} < \alpha 1_j$
$d_{n,j} > \alpha 2_j$
$snr_{n,j} < \alpha 4_j$

09/806756

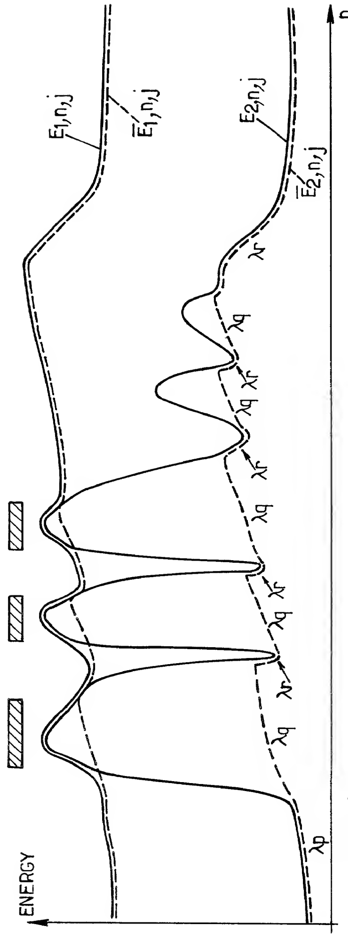


FIG.5.

09/806756

FIG.8

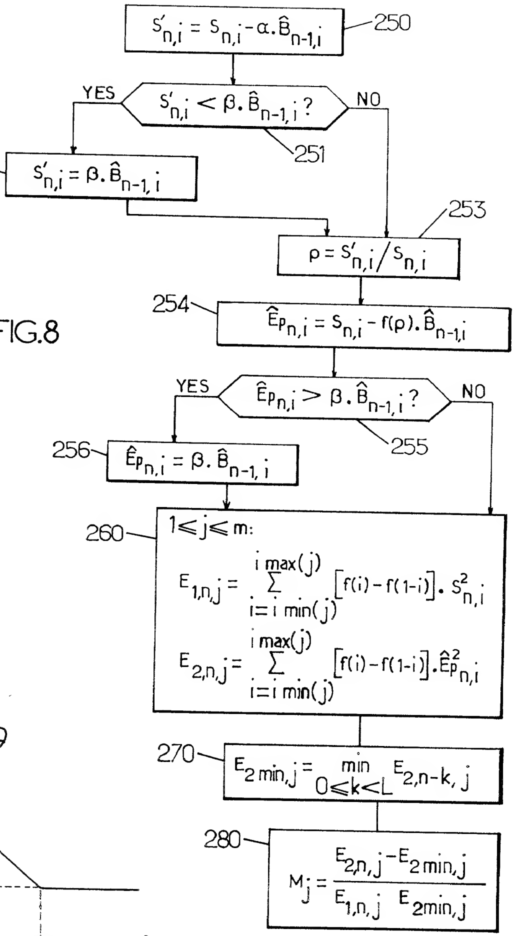
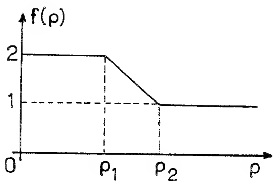


FIG.9



Attorney's Docket No.: US

PATENT

DECLARATION AND POWER OF ATTORNEY FOR PATENT APPLICATION

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below, next to my name.

I believe I am the original, first, and sole inventor (if only one name is listed below) or an original, first, and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled METHOD AND DEVICE FOR DETECTING VOICE ACTIVITY

the specification of which

X	is attached hereto.
X	was filed on 2 August 2000
	United States Application Number ________
	Or PCT International Application Number PCT/FR00/02220
	And was amended on ________ (if applicable)

I hereby state that I have reviewed and understand the contents of the above-identified specification, including the claim(s), as amended by any amendment referred to above. I do not know and do not believe that the claimed invention was ever known or used in the United States of America before my invention thereof, or patented or described in any printed publication in any country before my invention thereof or more than one year prior to this application, that the same was not in public use or on sale in the United States of America more than one year prior to this application, and that the invention has not been patented or made the subject of an inventor's certificate Issued before the date of this application in any country foreign to the United States of America on an application filed by me or my legal representatives or assigns more than twelve months (for a utility patent application) or six months (for a design patent application) prior to this application.

I acknowledge the duty to disclose all information known to me to be material to patentability as defined in Title 37, Code of Federal Regulations, Section 1.56.

I hereby claim foreign priority benefits under Title 35, United States Code, Section 119(a)-(d), of any foreign application(s) for patent or inventor's certificate listed below and have also identified below any foreign application for patent or inventor's certificate having a filing date before that of the application on which priority is claimed:

Prior Foreign Application(s):

Number	(Country)	(Day/Month/Year Filed)	Priority Claimed Yes	No
9910128	FRANCE	04/08/1999	X	

I hereby claim the benefit under title 35, United States Code, Section 119(e) of the United States provisional application(s) listed below:

(Application Number)	(Filing Date)

I hereby claim the benefit under Title 35, United States Code, Section 120 of any United States application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of Title 35, United States Code, Section 112, I acknowledge the duty to disclose all information known to me to be material to patentability as defined in Title 37, Code of Federal regulations, Section 1.56 which became available between the filing date of the prior application and the national or PCT International filing date of this application:

(Application Number)	Filing Date	(Status-patented, pending, abandoned)

I hereby appoint Timothy N. Trop, Reg. No. 28,994; Fred G. Pruner, Jr., Reg. No. 40,779, Dan C. Hu, Reg. No. 40,025 and Ruben S. Bains, Reg. No. 46,532; my patent attorneys, of TROP, PRUNER & HU, P.C., with offices located at 8554 Katy Freeway, Ste. 100, Houston, TX 77024, telephone (713) 468-8880, my patent attorneys; with full power of substitution and revocation, to prosecute this application and to transact all business in the Patent and Trademark Office connected herewith.

Send correspondence to Dan C. Hu, TROP, PRUNER & HU, P.C., 8554 Katy Freeway, Ste. 100, Houston, TX 77024 and direct telephone calls to Dan C. Hu, (713) 468-8880.

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Full Name of Sole/First Inventor: **Stéphane LUBIARZ**	
Inventor's Signature:	Date: 11/04/2001
Residence: **75016 PARIS (France)** FKX	Citizenship: FRENCH
Post Office Address: **4 Avenue Léon Heuzey, 75016 PARIS (France)**	
Full Name of Second/Joint Inventor: **Edouard HINARD**	
Inventor's Signature:	Date: 31/04/2001
Residence: **75015 PARIS (FRANCE)** FKX	Citizenship: FRENCH
Post Office Address: **26 Rue de la Fédératon, 75015 PARIS (France)**	
Full Name of Third/Joint Inventor: **François CAPMAN**	
Inventor's Signature:	Date: 01/03/2001
Residence: **78000 VERSAILLES (France)** FKX	Citizenship: FRENCH
Post Office Address: **47 rue des Etats Généraux, 78000 VERSAILLES (France)**	
Full Name of fourth/Joint Inventor: **Philip LOCKWOOD**	
Inventor's Signature:	Date: 01/03/2001
Residence: **95490 VAUREAL (France)** FKX	Citizenship: FRENCH
Post Office Address: 22 rue des Aulnes **95490 VAUREAL (France)**	

1-W

2-W

3-W

4-W